Guidelines for your BI project

1. Choose a dataset from one of the data repositories listed on the BI web page in the box named: **Use cases, datasets and readings .** The dataset must be sufficiently large and with a sufficiently large number of attributes. Analyze possible options, and select one for which you can think of a number of descriptive, predictive, prescriptive questions.
2. Check if the attributes in the dataset allow answering your business questions listed in step 1. It might be necessary to add new attributes (it must be possible to create these new "columns" out of existing ones, as for the example of *Total Revenue= Revenues-in-Europe + Revenues-outside-Europe*. It might also be necessary to structure le *values* of some attribute in hierarchies (e.g. date->week->month-> year; city->region->nation-> continent; etc). Take note of all these aspects.
3. Now you are ready to upload your data on Watson Analytics (WA). The first step is ETL (see slides *2.DataWharehousesETL* on the BI web page and the slides on Lab 2). First of all, check *data quality* (see bhow on Lab 2 slides). If below 60%, you may consider another dataset, or you can apply some data cleaning, as shown in Lab 2 slides. Rename attributes so that their meaning is clear. Next, apply *data trasformations* , as you already decided in step 2 of these guidelines. If necessary, merge attributes, create new attributes, group attribute *values* in hierarchies, or discretize attribute *values* if they are real numbers (e.g., age: >=50 or <50).
4. Start with **descriptive analytics**. Use the pre-defined natural language questions proposed by WA to explore and analyze your data. Some of these questions can also enrich your initial set of questions listed in step 1. Choose the appropriate visualization for showing the answers to these questions (see how on Lab 3 slides and lesson 6, DataVisualization). To answer your questions, WA will apply many of the operations described in lesson 3 on Data structures and processing, but fortunately, WA would do this for you automatically (slicing, dicing, rolling up and down, pivoting your dataset).
5. If you have textual data, you first need to transform these data. See slides on DataAnalytics and SocialAnalytics lessons, and last Labs (topic 6: work with unstructured data).
6. Now you are ready to use data analytic tools to answer **predictive** and **prescriptive** questions. Choose the **attribute values that you would like WA be able to predict**. If values are discrete (or discretized in step 3), you can use Decision Trees, if continuous, you can choose Regressions. For example, product categories are discrite (fixed set of values) revenues are continuous. See *DataAnalytics* lessons to learn the general ideas of decision trees and regression algorithms. This might be the most complex task. If you need one of my Machine Learning students to help you, send me an email. Don't worry if you can't complete this step perfectly, just do your best.
7. **Prescriptive** questions can be answered using results from predictive analytics. For example, say that you used the Decision Tree method to learn predictive rules. Say that the domain is e-commerce and one of the rules is something like: *IF AGE>30 and AGE<=45 and CITY= New York and LIKES= Country Music and (other preferences induced from users' previous purchase behaviors)* ***THEN*** *RECOMMEND= "Traveller" by Kane Brown* (Support 15%, Confidence 80%). This rule predicts that a user whose attributes match the left hand side of the rule (the part before the **THEN**), he/she is likely to buy the album "Traveller". If the rule has a sufficient support (many records in the database match the rule) and high confidence (probability of being a correct rule, given the examples in the dataset), then you may decide to use this rule to create a campaign to push sales of records by Kane Brown, and address the campaign to users with similar profiles. These issues are discussed in lessons about Data Analytics.
8. **Prepare** a report in which you describe your project and results, following all the steps of this guideline. Comment your discoveries. I am aware that steps 6 and 7 might be complex if you don't have a sufficiently "good" dataset, and that understanding if it is "good" or not is not easy for a non-IT expert. So, don't worry too much, just do your best. Steps 6 and 7, in principle, would demand for a cooperation with IT experts. But let's take the challenge!